We claim:

1.   A method for determining dominant phrase vectors in a topological vector space for a semantic content of a document on a computer system, the method comprising:

accessing dominant phrases for the document, the dominant phrases representing a
5   condensed content for the document;

constructing at least one state vector in the topological vector space for each dominant phrase using a dictionary and a basis; and

collecting the state vectors into the dominant phrase vectors for the document.

10   2.   A method according to claim 1, wherein accessing dominant phrases includes extracting the dominant phrases from the document using a phrase extractor.

3.   A method according to claim 1, wherein accessing dominant phrases includes storing the dominant phrases in computer memory accessible by the computer system.
15

4.   A method according to claim 1, the method further comprising forming a semantic abstract comprising the dominant phrase vectors.

5.   A method for determining dominant vectors in a topological vector space for a
20   semantic content of a document on a computer system, the method comprising:

storing the document in computer memory accessible by the computer system;

extracting words from at least a portion of the document;

constructing a state vector in the topological vector space for each word using a dictionary and a basis;
25   filtering the state vectors; and

collecting the filtered state vectors into the dominant vectors for the document.

6.   A method according to claim 5, wherein extracting words includes extracting words from the entire document.

30

7.   A method according to claim 5, wherein filtering the state vectors includes selecting the state vectors that occur with highest frequencies.

8.   A method according to claim 5, wherein filtering the state vectors includes:

calculating a centroid in the topological vector space for the state vectors; and

selecting the state vectors nearest the centroid.

9.   A method according to claim 5, the method further comprising forming a semantic abstract comprising the dominant vectors.

10.   A computer-readable medium containing a program to determine dominant vectors in a topological vector space for a semantic content of a document on a computer system, the program being executable on the computer system to implement the method of claim 5.

11.   A method for determining a semantic abstract in a topological vector space for a semantic content of a document on a computer system, the method comprising:

storing the document in computer memory accessible by the computer system;

determining dominant phrase vectors for the document;

determining dominant vectors for the document; and

generating the semantic abstract using the dominant phrase vectors and the dominant vectors.

12.   A method according to claim 11, wherein generating the semantic abstract includes reducing the dominant phrase vectors based on the dominant vectors.

13.   A method according to claim 11, wherein generating the semantic abstract includes reducing the dominant vectors based on the dominant phrase vectors.

14.   A method according to claim 11, wherein generating the semantic abstract includes obtaining a probability distribution function for a reduced set of the dominant phrase vectors similar to a probability distribution function for the dominant phrase vectors.

5   15.   A method according to claim 11, the method further comprising identifying the lexemes or lexeme phrases corresponding to state vectors in the semantic abstract.

16.   A computer-readable medium containing a program to determine a semantic abstract in a topological vector space for a semantic content of a document on a computer

10   system, the program being executable on the computer system to implement the method of claim 11.

17.   A method for comparing the semantic content of first and second documents on a computer system, the method comprising:

15   determining semantic abstracts for the first and second documents;
            measuring a distance between the semantic abstracts; and
            classifying how closely related the first and second documents are using the distance.

18.   A method according to claim 17, wherein measuring a distance includes

20   measuring a Hausdorff distance between the semantic abstracts.

19.   A method according to claim 17, wherein measuring a distance includes determining a centroid vector in the topological vector space for each semantic abstract.

25   20.   A method according to claim 19, wherein measuring a distance further includes measuring an angle between the centroid vectors.

21.   A method according to claim 19, wherein measuring a distance further includes measuring a Euclidean distance between the centroid vectors.

30

22.   A computer-readable medium containing a program to compare the semantic content of first and second documents on a computer system, the program being executable on the computer system to implement the method of claim 17.

5   23.   A method for locating a second document on a computer with a semantic content similar to a first document, the method comprising:

determining a semantic abstract for the first document;

locating a second document;

determining a semantic abstract for the second document;

10   measuring a distance between the semantic abstracts for the first and second documents;

classifying how closely related the first and second documents are using the distance; and

if the second document is classified as having a semantic content similar to the

15   semantic content of the first document, selecting the second document.

24.   A method according to claim 23, the method further comprising, if the second document is classified as not having a semantic content similar to the semantic content of the first document, rejecting the second document.

20

25.   An apparatus on a computer system to determine a semantic abstract in a topological vector space for a semantic content of a document stored on the computer system, the apparatus comprising:

a phrase extractor adapted to extract phrases from the document;

25   a state vector constructor adapted to construct at least one state vector in the topological vector space for each phrase extracted by the phrase extractor; and

collection means for collecting the state vectors into the semantic abstract for the document.

30   26.   An apparatus according to claim 25, the apparatus further comprising filter means for filtering the state vectors to reduce the size of the semantic abstract.

MJM Docket No. 6647-13
Novell IDR-432

27.   An apparatus according to claim 25, wherein the state vector constructor is further adapted to construct a state vector for each word in the document.

28.   An apparatus on a computer system to compare the semantic content of first and second documents on a computer system, the apparatus comprising:

first and second semantic abstracts for the first and second documents, respectively, stored on the computer system and represented as sets of vectors in a topological vector space;

measuring means for measuring the distance between the first and second semantic abstracts; and

a classification scale to determine how closely related the first and second documents are based on the distance between the first and second semantic abstracts.

29.   A method for determining a semantic abstract in a topological vector space for a semantic content of a document on a computer system, the method comprising:

extracting dominant phrases from the document using a phrase extractor, the dominant phrases representing a condensed content for the document;

constructing at least one first state vector in the topological vector space for each dominant phrase using a dictionary and a basis;

collecting the first state vectors into dominant phrase vectors for the document;

extracting words from at least a portion of the document;

constructing a second state vector in the topological vector space for each word using the dictionary and the basis;

filtering the second state vectors;

collecting the filtered second state vectors into dominant vectors for the document; and

generating the semantic abstract using the dominant phrase vectors and the dominant vectors.

30.   A method according to claim 29, the method further comprising comparing the semantic abstract with a second semantic abstract for a second document to determine how closely related the contents of the documents are.